International Journal of Educational Technology and Learning
ISSN: 2523-0581
Vol. 3, No. 1, pp. 17-23, 2018
DOI: 10.20448/2003.31.17.23

A Review of Educational Data Mining Tools & Techniques

Haitham Alagib Alsuddig Hamza[1]
Piet Kommers[2]

[1]College of Computer Science and Information Technology, Sudan University of Science and Technology, Sudan.
Email: haithamalagib@hotmail.com
[2]Faculty of Behavioural Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands.

Abstract

The purpose of this paper is to present a review study on the use of the tools and techniques of Education Data Mining (EDM); through the use of data mining techniques such as classification, aggregation, the correlation rules and the disclosure of cases and apply on the student grades to show the benefits of applying the techniques of data mining in the academic field to obtain a clear understanding of the factors success and failure of students. The paper presented a definition of the concept of (EDM) and then presents of several previous studies that used many different techniques of data mining. The paper concluded to review the most important applications of (EDM) and classification according to the objectives and methods used and General framework procedural can be use in (EDM).

Keywords:
Educational data mining
Decision trees
performance
student
Predictive.

Licensed:
This work is licensed under a Creative Commons Attribution 4.0 License.

Publisher:
Scientific Publishing Institute

1. Introduction

Education Data Mining (EDM) is growing at a very fast pace. The main aim of EDM is to develop methods to explore the unique type of data that comes from educational institutes and to use those methods to understand the students and their learning environments. EDM deals with mining of large data sets of educational data to answer educational research questions. These data sets may come from learning management systems, interactive learning environments, intelligent tutoring systems, or any system used in a learning context.

The educational institutions management process is one of the difficulties faced by the supervisors because of the large size and complexity of its structure and the multiple sources of data. Therefore, the educational institution faces several problems during the management of the educational process, including academic, financial and administrative. These problems need to be studied, conclusions and recommendations that contribute to the decision-making process that facilitates the process of education based on an information system that is built in advance in a modern scientific way. In the data of the institutions on students, graduates and faculty members, and the correlation of all this with the most important indicators of performance such as student achievement, survival rate or leakage and quality of performance of faculty members.

One of the biggest problems affecting in the performance of the educational process in universities is Student cannot decide about their field of study before they are enrolled in specific field of study. The research problem is to find a method for making student choices more explicit and more fitting to their capacities and motivation.

Higher education institutions are beginning to use analysis to improve services they provide and for increasing student upgrade and retention. The U.S. National Department of Education and Technology Plan, as one part of its model for 21stcentury learning powered by technology, envisions ways of using data from online learning systems to improve instruction (Mining, 2012).

The purpose of this paper to know some of the methodologies and tools & Techniques of EDM which have been used in previous studies to find the best of them and evaluate the performance of the algorithms for classification and exploration of data to obtain the highest accuracy and the lowest proportion of the line when used in mining and build the models.

2. Literature Review

Many studies around the world have been interested in applying data mining algorithms to discover knowledge in universities ,One of the most important of these studies is a study concerned with the

applications of data mining in the field of higher education, Focused on the input of the educational process and its outputs and how they affect each other, The study used the method of neural networks to explore data, The results showed diverse relationships between curricula , multiple hours , the nature of students and between the graduates and the jobs they occupy, As well as other useful conclusions for decision-makers at universities.

Another study focused on prospecting at the City University site by studying guest registration files, the study concluded many relationships between the nature of visitors and the nature of the disciplines they study, in addition to the reasons for the infiltration, of customers from the institution, and made many recommendations to decision-makers at the university (Lekeas, 2000).

Another study presents an applied study in data mining and knowledge discovery. It aims at discovering patterns within historical students' academic and financial data at UST (University of Science and Technology) from the year 1993 to 2005 in order to contribute improving academic performance at UST. Results show that these rules concentrate on three main issues, students' academic achievements (successes and failures), students' drop out, and students' financial behaviour. Clustering (by K-means algorithm), association rules (by Apriori algorithm) and decision trees by (J48 and Id3 algorithms) techniques have been used to build the data model. Results have been discussed and analyses comprehensively and then well evaluated by experts in terms of some criteria such as validity, reality, utility, and originality. In addition, practical evaluation using SQL queries have been applied to test the accuracy of produced model (rules), the shortcomings of this study are as follows:

a) Not included of Associate student data, educational staff and other university branches.
b) Not included in scholarship data and lots of personal data for students.
c) Not included attendance and absence data for students (Al-shargabi, 2010).

Another study was conducted in Ethiopia In Debre_Markos University study has shown that data mining techniques can be applied by higher education institutions or universities in determining student failure/success rate so that managing students' enrolment at the beginning of the year, assist students before they reached risk of failure, effective resource utilization and cost minimization, helping and guiding administrative officers to be successful in management and decision making. The study applied data mining technology to the data of university students for the purpose of forecasting the success or failure of students, the study used CRISP methodology the analysis was carried out by the WEKA program and the forecast model was built the study found the main class, number of courses given in a semester, and field of study are the major factors affecting the student performances (Asif, Merceron, Ali, & Haider, 2017).

One of the studies discussed that Student performance in university courses is of great concern to the higher education managements where several factors may affect the performance. This study is an attempt to use the data mining processes, particularly classification, to help in enhancing the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses (Gulati, 2012).

Ramaswamy N has written research paper which focused on predicts student's characteristics or academic performances in various educational institutions. This paper focus on students' performance as a slow learner or fast Lerner. For that they applied various data mining techniques and compare the accuracy based on student's attributes. For assessing the goodness of a predictor, an extensive study on the student data set was conducted by applying five individual classifiers J48 (J48), Bayesian Net (BN), Neural Net (NN), Decision Tree (DT), and Naïve Bayes (NB) (Ramaswami, 2014).

The other research paper which is written by Mrs. M.S. Mythili, Dr. A.R. Mohamed Shanavas to use data mining methodologies to study and analyses the school students' performance based on classification techniques which is useful to gauge students' performance and deals with the accuracy, confusion matrices and the execution time taken by the various classification data mining algorithms. The decision tree classifier C4.5 (J48), Random Forest, Neural Network (Multilayer Perception) and Lazy based classifier (IB1) Rule based classifier (Decision Table) were enforced in weak (Mythili, 2014).

Using K-means clustering which used for pattern recognized classification and clustered students according to their class performance, sessional and attendance record. Using K-Means Clustering clustered the students based on their Class Performance, sessional and Attendance in class. Centroids are calculated from the educational data set taking Kclusters. This study is helpful to notify the students with less attendance and slow performance in sessional but also enhances the decision-making approach to monitor the performance of students. Also, on increasing the value of K, the accuracy becomes better with huge dataset and Kmeans can find the better grouping of the data. The results obtained help to cluster those students who need special attention (Guleria, 2014).

One of researches have conducted comparison of data mining algorithms for clustering published. These algorithms are among the most influential data mining algorithms in the research community. A Knn algorithm is more sophisticated approach, k-nearest neighbour (kNN) classification, finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighbourhood. KNN classification is an easy to understand and easy to implement classification technique. Despite its simplicity, it has done perform well in many situations (Kushwah, n.d.).

Another study explored the opportunities of the Education data mining for improving students' performance. Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it. The study used Classification methods like decision trees, Bayesian network etc. which can be applied on the educational data for predicting the student's performance in examination. This prediction will help to identify the weak students and help them to score better marks. The C4.5, ID3 and CART decision tree algorithms are applied on engineering student's data to predict their performance in the final exam. The results of this study provided predicted the number of students who are likely to pass, fail or promoted to next year, steps to improve the performance of the students who were predicted to fail or promoted and the comparative analysis of the results states that the prediction has helped the weaker students to improve and brought out betterment in the result (Yadav, 2012).

One of the studies showed how useful data mining can be in higher education in particularly to improve student performance through educational data mining to analyses learning behaviour. This study collected students' data from Database Course After pre-processing and applied data mining techniques to discover association, classification, clustering and outlier detection rules, in each of these four tasks, the study extracted knowledge that describes students' behaviour, Also, experiments could be done using more data mining techniques such as neural nets, genetic algorithms, k-nearest neighbors, Naive Bayes, support vector machines and others. The study also recommended for used pre-process and data mining algorithms could be embedded into eLearning system so that anyone using the system can benefited from the data mining techniques (El-Halees, 2009).

One of the studies discussed suggests that legal access to alcohol does affect student performance. The study concluded to prediction of teenager's alcohol addiction by using demographic, family and other data related to student, different classifiers are studied and the experiments are conducted to find the best classifier for predicting the performance of the students who consume alcohol. The study propose an approach to predict the performance using data mining techniques, the study also shows that the most important attributes which most affected the performance of students who consume the alcohol during their study are the previous grades which is gained by students and other attributes are absence in the class, father's job, mother's job, extra educational support, extra paid classes within the course subject, wants to take higher education, reason to choose this institution and also some other attributes (Pal, 2017).

Another study looks at and compare well performing algorithms such as Naïve Bayes, decision tree (J48), Random Forest, Naïve Bayes Multiple Nominal, K-star and IBk. And it mentions Educational Data mining is a relatively new field and has a lot of potential to help society if used in the proper manner. The study compared six algorithms J48 (Decision Tree), Random Forest, Naive Bayes, Naive Bayes Multinomial, K-star, IBk. In the comparative study of all these algorithms can see that the closest we got in terms of getting an accurate prediction was the Random Forest Technique which narrowly edged the J48 to claim the top spot. This was that was done on a relatively larger dataset hence random forest becomes more accurate with the number of entries but all algorithms need modification if they can ever be used because the current amount of accuracy is low for this to be implemented on a large scale in the present state .

Another research considered that one of the common tools to evaluate instructors' performance is the course evaluation questionnaire to evaluate based on students' perception. In this study, classification algorithm of Naïve Bayes and C5.0 are used to build classifier models. Their performances are compared over a dataset composed of answer of students to a real course evaluation questionnaire using accuracy, precision, recall, and specificity performance metrics. Although all the classifier models show comparably high classification performances, Naïve Bayes classifier is the best with respect to accuracy, precision, and specificity. In addition, an analysis of the variable importance for each classifier model is done. This research describes the performances of classification algorithms used in building a model does not necessarily indicate that the one that used the least time is the best model to use. Some Algorithms can take the least time but may not produce the best result in term of accuracy. This research used classification algorithms and data mining techniques such as, Naïve Bayes classifier, C5.0 as well as data from universities. Naïve Bayes classifier is the best with respect to accuracy, precision, and specificity (Patil, 2017).

Another research used data mining techniques for predicting the students' graduation performance in final year at university using only pre-university marks and examination marks of early years at university, no socio-economic or demographic features are use. The result of the study shows that can predict the graduation performance in a four-years university program using only pre-university marks and marks of first and second year courses, no socio-economic or demographic features, with a reasonable accuracy, and that the model established for one cohort generalizes to the following cohort. It makes the implementation of a performance support system in a university simpler because from an administrative point of view, it is easier to gather marks of students than their socio-economic data. The result also shows that decision trees can be used to identify the courses that act as indicator of low performance. By identifying these courses can give warning to students earlier in the degree program (Asif et al., 2017).

This paper, presented a comparative study on the effectiveness of educational data mining techniques to early predict students likely to fail in introductory programming courses. Although several works have analysed these techniques to identify students' academic failures. The study evaluated the effectiveness of four

prediction techniques on two different and independent data sources on introductory programming courses available from a Brazilian Public University: one comes from distance education and the other from on-campus. The results showed that the techniques analyses in this study are able to early identify students likely to fail, the effectiveness of some of these techniques is improved after applying the data pre-processing and/or algorithms fine-tuning, and the support vector machine technique outperforms the other ones in a statistically significant way.

In a one of the studies, the researchers used ID3 algorithm to classify students and predict what causes student failure so that teacher can help them to avoid failure (Baradwaj, 2012).

In another study, the researcher used a wide range of tree resolution algorithms rule-learners, J48, Bayes, Naiva, Bayes and IBK. The researcher concluded that the algorithm of J48 is the most accurate algorithm in terms of the accuracy of the prediction, noting that the accuracy of the prediction of the previous algorithms in general was not satisfactory and it cannot be relied on it, where the accuracy of the prediction was promoted between (67% -52%). And use more data to improve the data collection process (Kabakchieva, 2013).

In further study, the researchers predicted the performance of students based on their academic levels in several areas; using the detection of the rules of correlation to ensure that the factors affecting the final outcome of the student linked to each other, having calculate the correlation coefficient while the student attributes were shown and the result was different from the resulting relationship using the detection of the rules of correlation (Borkar, 2013).

In this cited study below, the researchers used three techniques for Data Mining: classification, clustering and detection of the rules of association on the data collected by students from an e-learning system. This study is a theoretical and practical guide to how to apply data mining techniques to the educational system (Romero, 2008).

In another study, researchers apply a classification ID3, J48 and Apriori algorithm to reveal the correlation rules in the system data; Moodle to compare the accuracy of the algorithms in terms of their ability to predict the outcome of the student, whereas they conclude the algorithm ID3 is more accurate than the rest of the algorithms with a probability of 83.916% (Kularbphettong, 2012).

3. Results

The spread of the use of educational information systems in the institutions of higher education and the emergence of new concepts in teaching and learning, such as e-learning and distance learning to the availability of a large amount of data mining from these systems, which led to the search for ways to extract information to improve the performance of students and teachers. There are different methods of data mining depending on the types of data applied, as data patterns can be classified into the following basic categories:

Table-1. Patterns of data extracted from educational information systems.

Data Pattern	Methods used in data mining
Structured organizational data extracted from relational databases	Basic exploration methods are used, such as classification, clustering, correlation and prediction relationships
Historical data expressed in time series	Special prospecting methods applied to time series such as prediction, Expectation and correlation study
Texts	Methods of textual exploration
Multimedia data such as images, audio and video	Data mining methods of multimedia
Data generated by web applications	Data manipulation methods provided by Web applications are three different forms: 1 - Prospecting in the content of pages. 2. Prospecting in the structure of pages. 3-Excavation in the records of the course.

The methods used in educational data mining, they are same methods used in the traditional methods of data, the need to understand the environment that will be handled and collect data to clean, arrange and select the techniques to be applied and finally interpretation of the results as well as verify the validity of the techniques being applied, taking into account the different methods; Using results from the privacy of educational environment and purpose of prospecting. The procedure of mining in educational data can be summarized as follows:

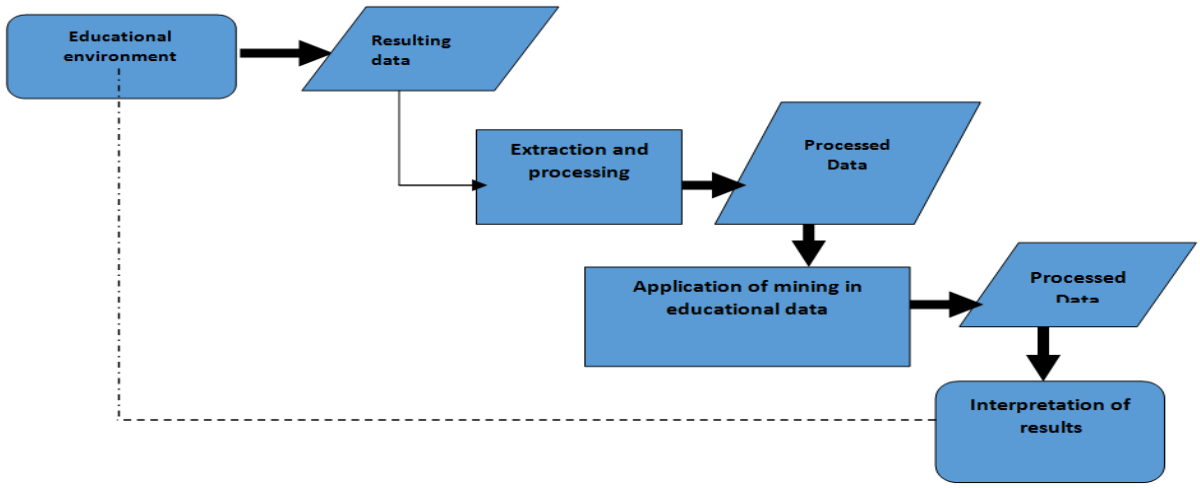

Figure-1. General framework procedural in Educational Data Mining.

There are many methods are used in the process of educational data mining in according to the application and the objectives for which used; the following is review of the most important of these general methods in addition to the most important applications:

Table-2. The common methods between the educational data mining and analysis of the learning process.

Methods	Objectives	Basic applications
Prediction	Predict the values of variables based on the values of other variables being filtered. Forecasting uses the following basic methods: decision tree, classifications, aggregation, and neural networks	There are many applications to be used to predict a student result, and accordingly correcting student behaviour to get the best expected performance.
Clustering	Assemble students in homogeneous and similar groups	Identify appropriate mechanisms to deal with similar student groups in learning style and social communication method.
correlations Relationship	Find causal relationships between variables and the most important methods in use.	Discover the weaknesses of the learners to improve them. Studying causal relationships in the educational process and discovering patterns of weakness to improve them.
Text Mining	Extract valuable information from text.	Analyse student conversations in forums to discover problems. Analysing the file of the movements resulting from the student wandering in the educational system in order to follow him and extract useful information about his interests.
Social Network Analysis	Discover and analyse relationships through social networks	Analysing the nature of relationships and interaction in communication networks; interactive tools in order to discover the student's educational style and discover weaknesses and preferences and other applications.

Table-3. The most important applications of educational data mining according to beneficiary.

Beneficiary	Applications of Educational Data Mining
Student	To discover the weaknesses of the student and suggest educational resources and educational activities help in improving his level. To discover the student's learning style in order allocates a learning session for each student.
Teacher	To help to obtain objective analysis and feedback on the method of education in order to improve it. To identify students who are in need of support. To expect student performance for direction guidance. To categorize students according to their levels or educational. methods. To identify activities of the most active students in delivering knowledge. To improve the allocation of educational content.
Designers of methods, programs and study plans	To evaluate and improve curricula in terms of content. To evaluate and improve study plans. To identify the teacher's educational model; and the student's model as well as design the study programs accordingly.
Higher administration of educational institutions	To Improve the decision-making process at the level of senior management after studying the indicators resulting from the use of mining methods
Managers of educational systems	To determine the best way to display and design electronic educational content. To choose the best design for distance learning. To Identify and preserve the value of indicators that should be studied to improve the quality of education. Best artistic design.

These studies show the values of the algorithms, the data sample, the aggregate and the characteristics of the students as factors that affect the accuracy of the algorithms. The previous studies show that; the Educational Data Mining follows the same approach as traditional methods of data mining, this Methodology begins with business understanding, then captures and understands data, data preparation and applies data mining techniques for building model, evaluates the model results, and deploys the knowledge gained in operations.

This paper finding that there are several different methods of data mining, and the choice of the appropriate method depends on the nature of the data under study and its size like analysis Correlation, decision tree, genetic algorithms, virtual theory networks, raw group path, neural network, statistical analysis and There are several tools to explore the data, the most important of which are Summarization, Classification, Prediction, clustering, Rule Analysis, and change and deviation detection.

Through the study I found that the most appropriate tools used in the educational data mining is the prediction tool, some of the traditional tools used in Predictions are, for example, regression and differential analysis. The new methods include correlation rules, decision tree, neural networks, and genetic algorithms.

4. Conclusions

Significance of this paper by considering the results of this study is expected to fill knowledge gap of implementation of tools and Techniques of data mining which help in building a model can be applied to increase the performance of the high education through mining in educational information systems which provide us with different types of data that can be applied to extracted the knowledge. The method of exploration varies depending on the types of data that can be applied. Also, data mining systems use mathematical, statistical and intelligent methods for building future forecasts and exploring behaviour and trends, allowing for estimation of decisions Correct and take them in time. The core of decision support systems is data mining and prediction, early warning and scenario formulation based on simulation models, where decision support systems synthesize available data with personal visions of the decision maker, is done across a range of mathematical models to predict and simulation.

References

Al-shargabi. (2010). Discovering vital patterns from UST students data by applying data mining techniques.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177-194.

Baradwaj, B. K. (2012). Mining educational data to analyse students' performance. arXiv preprint arXiv:1201.3417, 2.

Borkar, S. A. (2013). Predicting students' academic performance using education data mining. *International Journal of Computer Science and Mobile Computing, 2*, 273--279.

El-Halees, A. (2009). Mining students data to analyse e-Learning behaviour: A case study. el2009mining.

Gulati, P. A. (2012). Educational data mining for improving educational quality. *tell us*, 3.

Guleria, P. A. (2014). Mining educational data using K-means clustering. *guleria2014mining*.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *kabakchieva2013predicting*, *13*, 61--72.

Kularbphettong, K. A. (2012). Mining educational data to analyse the student motivation behavior. *World Academy of Science, Engineering and Technology*, *6*, 1036--1040.

Lekeas, G. K. (2000). Data mining the web: the case of City University's Log Files. *lekeas2000data*.

Mining, T. E. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Paper presented at the Proceedings of Conference on Advanced Technology for Education.

Mythili, D. A. (2014). An analysis of students' performance using classification algorithms. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 16.

Pal, S. A. (2017). Performance analysis of students consuming alcohol using data mining techniques. *International Journal of Advance Research in Science \& Engineering*, *6*, 238--250.

Patil, P. S. (2017). Predicting instructor performance using na{\"\i}ve bayes classification algorithm in data mining technique: A survey. *International Journal of Advanced Electronics and Communication Systems*, 6.

Ramaswami, M. (2014). Validating predictive performance of classifier models for multiclass problem in educational data mining. *International Journal of Computer Science Issues (IJCSI)*, *11*(86).

Romero, C. A. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*, 368--384.

Yadav, S. K. (2012). Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv:1203.3832.